D690–D697 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 25 November 2008*
doi:10.1093/nar/gkn828

Ensembl 2009

T. J. P. Hubbard[1,*], B. L. Aken[1], S. Ayling[1], B. Ballester[2], K. Beal[2], E. Bragin[1], S. Brent[1], Y. Chen[2], P. Clapham[1], L. Clarke[1], G. Coates[1], S. Fairley[1], S. Fitzgerald[2], J. Fernandez-Banet[1], L. Gordon[2], S. Graf[2], S. Haider[2], M. Hammond[2], R. Holland[2], K. Howe[1], A. Jenkinson[2], N. Johnson[2], A. Kahari[2], D. Keefe[2], S. Keenan[2], R. Kinsella[2], F. Kokocinski[1], E. Kulesha[2], D. Lawson[2], I. Longden[2], K. Megy[2], P. Meidl[2], B. Overduin[2], A. Parker[1], B. Pritchard[1], D. Rios[2], M. Schuster[2], G. Slater[2], D. Smedley[2], W. Spooner[2], G. Spudich[2], S. Trevanion[1], A. Vilella[2], J. Vogel[1], S. White[1], S. Wilder[2], A. Zadissa[1], E. Birney[2], F. Cunningham[2], V. Curwen[1], R. Durbin[1], X. M. Fernandez-Suarez[2], J. Herrero[2], A. Kasprzyk[2], G. Proctor[2], J. Smith[1], S. Searle[1] and P. Flicek[2]

[1]Wellcome Trust Sanger Institute and [2]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received and Accepted October 14, 2008

ABSTRACT

The Ensembl project (http://www.ensembl.org) is a comprehensive genome information system featuring an integrated set of genome annotation, databases, and other information for chordate, selected model organism and disease vector genomes. As of release 51 (November 2008), Ensembl fully supports 45 species, and three additional species have preliminary support. New species in the past year include orangutan and six additional low coverage mammalian genomes. Major additions and improvements to Ensembl since our previous report include a major redesign of our website; generation of multiple genome alignments and ancestral sequences using the new Enredo-Pecan-Ortheus pipeline and development of our software infrastructure, particularly to support the Ensembl Genomes project (http://www.ensemblgenomes.org/).

INTRODUCTION

The genome sequence of an organism provides a natural index for organizing and understanding biological data. The Ensembl project provides a comprehensive genome information system consisting of data storage, integration, analysis and visualization of a wide variety of biological data. Ensembl's primary focus is around providing gene annotation and comparative genome integration for chordate genomes, the vast majority of which are vertebrates. Ensembl concentrates particularly on mammalian genomes having developed initially around the human genome sequence. In comparison to similar projects based at the University of California Santa Cruz (1) and the National Center for Biotechnology Information (2), some of the distinguishing characteristics of the Ensembl project are:

(1) It provides consistent sets of annotation data within and between genomes:
   - It provides a geneset for each genome, generated from an automatic pipeline where no manually curated geneset exists, with stable identifiers which are tracked between Ensembl releases.
   - It provides relationships between genes and genomes in a comparative genomics framework in the form of sequence alignments, ortholog and paralog assignments and genetrees, again generated from an automatic pipeline where no manually curated relationships exist.

(2) It is a completely open project, not only through providing downloads of all data and software source code, but through multiple levels of programmatic access:
   - It allows its database system to be programmed against using the Ensembl API, a powerful object

*To whom correspondence should be addressed. Tel: +44 1223 496886; Fax: +44 1223 494919; Email: th@sanger.ac.uk

oriented representation of biological entities (e.g. genes) implemented in the Perl programming language.
- It allows its genome annotations, alignments, variation and functional genomics data to be dynamically federated with external data sources via the DAS protocol (3,4) and visualized through many of its website interfaces (http://www.ensembl.org/).
- It allows its datasets to be dynamically federated with external datasets for data mining using the BioMart system (5).

The Ensembl project is now being joined by the Ensembl Genomes project (http://www.ensemblgenomes.org/), which will use Ensembl technology to ultimately provide a common interface to genomes across biology.

A continuing driver for developments in Ensembl is its active involvement in many data generation and analysis projects. Recent examples have been the Rat haplotype project (6–8) and the ENCODE project (9). Dealing with data generated by the ENCODE project in particular had led to the development of specific algorithms for experimental data handling, such as approaches for designing and assessing whole genome tiling arrays (10). Ensembl has continued to be strongly involved in analysis for publications of new vertebrate genome sequences, particularly through its genesets (11–13) (see below).

The report lists only some of the new features, new data and other improvements that we have added to Ensembl since our last report (14). Users interested in the most up-to-date details of the Ensembl project should visit the Ensembl main page (http://www.ensembl.org) and follow the 'What's new' link and/or subscribe to the low-volume 'Ensembl announce' mailing list by sending email 'subscribe ensembl-announce' as the message body to majordomo@ebi.ac.uk. There is also an Ensembl blog (http://ensembl.blogspot.com/) and associated RSS feeds which in particular cover upcoming Ensembl training courses around the world (see below). Users with questions about Ensembl can consult the extensive online help, FAQ and tutorial materials (15) (include animated tutorials) or contact the Ensembl helpdesk through the website or by emailing helpdesk@ensembl.org.

RESULTS

Ensembl Web site redesign

The majority of users access Ensembl through its web interface, making it a critical component of the project. It is generally recognized that major factors influencing website usability are speed and discoverability. As websites grow and their underlying databases become more complex, individual web pages tend to become larger, more complex and slower to display and it becomes harder for users to discover new functionality and navigate to the pages most appropriate to their query. The case of Ensembl is no different: the data contained in its databases is presented in different ways through a number of different 'views', which have been progressively added as project has developed, starting with relatively straightforward views summarizing information about a given gene, or displaying a region of genome sequence (16), to increasingly complex views such as TranscriptSNPView (17) showing sequence variation within a given transcript across a set of strains or individuals. At the same time, the amount of data contained in many views has grown, for example the increased number of species has greatly added to the data presented in views containing comparative genome information. It is not straightforward to identify bottlenecks for users in web based systems. For example, analysing web log files does not easily distinguish between web pages which are of interest to a limited number of users and pages which most users have not discovered. Perceived web site performance can also be very different for different users as a result of different browsers, desktop machines and network speeds. Since the last report (14) considerable effort has been invested in understanding and addressing these issues, culminating in a substantially redesigned and reengineered website from Ensembl release 51 (November 2008).

In the new design (release 51) the web-code has been completely re-developed with improved speed as a high priority. The changes result in substantially smaller web pages which load much faster. A single page now requires far fewer network connections to the web servers, which substantially improves performance for users distant from the web servers. This has been achieved through the adoption of standards compliant HTML, Javascript and CSS; a more streamlined use of the AJAX (Asynchronous JavaScript and XML) protocol to include additional content; the incorporation of shared memory caching using memcached (http://www.danga.com/memcached/); and optimized Apache web server settings to improve browser performance. To enable the project to prioritize improvements and measure their impact on speed, a system for continuous automated monitoring of the response speed of the Ensembl website from more than ten sites around the world was developed and deployed in early 2008.

In parallel with the redevelopment of the underlying web-code, the website has been redesigned to improve navigability and discoverability (Figure 1). The new design organizes different views into four classes: Location, Gene, Transcript and Variation, which can be easily navigated between through tabs at the top of each web page. The location class includes views of the genome sequence at a range of resolutions and genome sequence based comparative views (Figure 1A). Gene based views include textual information about the gene, views of its local genomic environment, views of the gene in the context of its orthologs and paralog relationships with other genomes in the Ensembl system and views of sequence variation within that population (Figure 1B). Transcript based views are similar to the gene based ones, but focus around individual transcript structures with more detail (Figure 1C). Variation based views display information focused around individual SNPs (data not shown). Information presented in a single view in previous versions of Ensembl is now presented as separate smaller views in the new design. The relationship between these new views is clearly shown by the left hand hierarchical

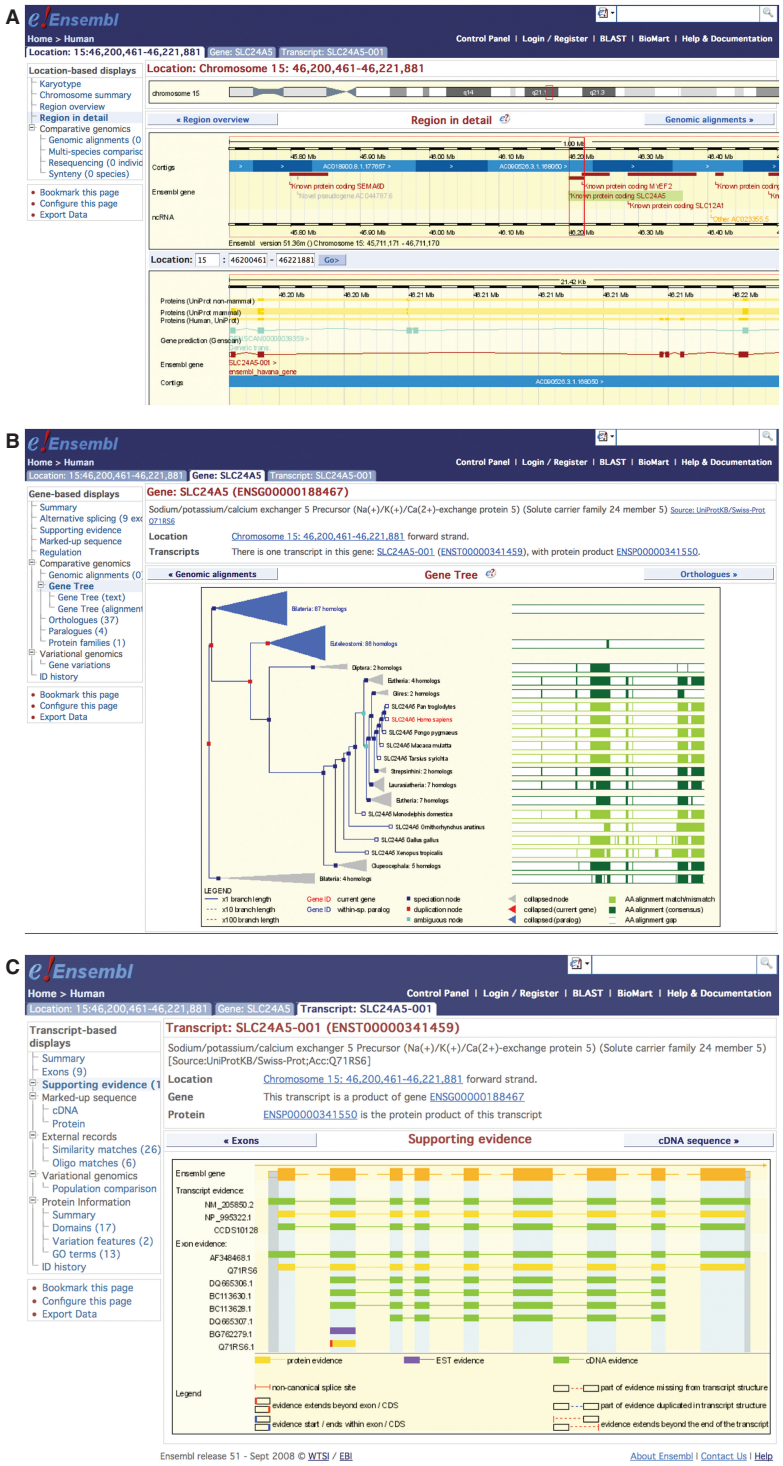

Figure 1. Screenshots of the Release 51 Ensembl website illustrating the principles of the new design and some of the new features. The figure shows an example of three of the four classes of display view using human gene SLC24A5 as the context. (**A**) An example of a location based view, showing a region of the genome around the gene. (**B**) An example of a gene based view, showing the gene tree. (**C**) An example of a transcript based view, showing supporting evidence for the transcript model. The three tabs across the top of the page, allow rapid navigation between the three classes of view. The fourth variation tab (data not shown) appears if an individual SNP is selected. For each class, the left hand menu lists the different views available. For the location-based views (A), this includes views of a genome at a range of resolutions and genome sequence based comparative genomic views. For the gene-based views (B), this includes textual information about the gene, views of its local genomic environment, views of the gene in the context of its orthologs and paralog relationships with other genomes in the Ensembl system and views of sequence variation within that population. The transcript based views (C) have views similar to the gene based views, but focused around individual transcript structures with more detail. As well as the overall redesign of the navigation between views, there are substantial improvements to many individual views, based on the much more extensive use of AJAX in the new web-code. Examples are the genetree view (B) which allows nodes to be expanded or collapsed interactively, making the view much more usable for large gene families; the substantially redesigned supporting evidence views (C) and the page configuration options on many views (e.g. A) which are much more intuitive than before and have a much greater range of display options.

menus which is context specific for each class. Each view within a class has a common header panel, summarizing the location or object. Clear and easy navigation between views is provided through the left hand menu and the left and right buttons below the header panel. Since only a specific chunk of information in shown in each view, this makes pages easier to read as well as improving the responsiveness of the servers. Configuration controls have been considerably improved and now take the form of a context specific pop-up panel for most views, e.g. allowing tracks to be enabled and disabled in genome sequence based display elements. The same panel contains controls to allow external data to be uploaded into Ensembl, or for external data sources to be federated (DAS).

The ideas for the new design were developed and tested through extensive interactions with users, including one to one sessions, testing sessions of design mock ups and web-based questionnaires. Questions investigated preferences between alternative overall layouts (e.g. use of tabs/left hand menu bars) as well as detailed behaviour such as the preference for a consistent name for the protein product of transcript (translation, peptide, protein). The results of these surveys have led to a design which is user driven and was significantly different from the one we had initially planned. We will be maintaining a user panel to help in guiding interface development.

New species and improved gene annotations

In the past year, seven new species (all mammals) were added to Ensembl including one new high coverage genome *Pongo pygmaeus abelii* (orangutan) and six new low coverage genomes [*Pteropus vampyrus* (megabat), *Tursiops truncates* (dolphin), *Tarsius syrichta* (philippine tarsier), *Lama pacos* (alpaca), *Dipodomys ordii* (kangaroo rat) and *Procavia capensis* (rock hyrax)]. Ensembl now supports 19 low coverage 2× genome sequences, the majority generated as part of the Mamalian Genome Project (http://www.broad.mit.edu/node/296). So far only one of the original 2× genomes, *Cavia porcellus* (Guinea Pig), has been upgraded to high coverage (6.8×). Together with the other 13 high coverage mammalian genomes, Ensembl contains a total of 32 mammals, making it an extensive resource for mammalian comparative genomics. In total Ensembl now supports 48 genomes, 41 of which are vertebrates.

One of the major goals of Ensembl is to provide genesets which are as accurate and complete as possible and these continue to be used as reference genesets in analysis of new vertebrate genomes. Recent genome publications based on Ensembl genesets include those of Platypus *Ornithorhynchus anatinus* (11), the Oposum *Monodelphis domestica* (12) and the Rhesus Macaque *Macaca mulatta* (13). The gene build process is based on alignments of protein and cDNA sequences and there is continuous work to improve it and generate updated, more accurate and complete genesets. Different gene build strategies are used depending on the assembly, quality of the genome, its distance to high quality genomes and the extent of its organism-specific transcript evidence as has been previously described (18). This year one focus has been to develop a systematic post gene build comparative analysis process (using the Ensembl compara homology pipeline) to identify initial gene structures that appear to be evolutionarily inconsistent. These regions are then subject to a second, more computationally expensive localized gene build pipeline with more sensitive parameters. The major classes of problems identified are split genes, missing orthologous genes, partially predicted genes and false exons. For the test case of the horse genome with initially 20 322 gene models, this post-processing pipeline identified 236 genes that were split; added 1013 genes that had initially been missed, but for which there were orthologs; extended 1330 partially predicted genes and removed 840 false exons. The process is now being systematically applied to other high coverage mammalian genomes. These genesets will be patched in subsequent Ensembl releases.

The other major focus has been the ongoing improvement of the human geneset in collaboration with other groups. Ensembl, together with the Sanger Institute HAVANA group (19), is part of multiple collaborations to refine the human geneset including the CCDS (Consensus Coding Sequence) consortium, with RefSeq at NCBI (20) and UCSC (1), and the new ENCODE scale-up project GENCODE (http://www.sanger.ac.uk/encode/) with multiple collaborators. CCDS (http://www.ncbi.nlm.nih.gov/CCDS/) is a stable set of protein coding gene structures for which all consortium members agree to the base pair. Since our previous report (14) the human CCDS set has increased from 18 290 to 20 159 CDSs, which represents an increase from 16 003 to 17 052 genes with at least one CCDS entry. There is also a CCDS set for mouse, which has increased even more, from 13 374 to 17 707 CDSs and from 13 014 to 16 889 genes. GENCODE builds on CCDS to validate additional transcripts and extend into UTR regions, building on the ENCODE pilot project (9,21–23) and incorporating additional computational and experimental input and validation (24). One new computational approach, which is being built on within GENCODE, is to use alignments across the many mammalian genomes now available to evaluate the conservation of putative coding sequences (25). Several hundred transcript predictions generated by the Ensembl gene build pipeline which were found to have low scores in this analysis have been identified as spurious and are now filtered out. The Ensembl/HAVANA collaboration includes further efforts to improve geneset consistency, such as tighter links with UniProt (26) and input into the Genome Reference Consortium (http://www.sanger.ac.uk/sequencing/grc/) to flag discrepancies between the human genome sequence and transcript evidence.

The Ensembl/HAVANA human geneset shown in Ensembl is a combined output from these projects, incorporating all CCDS entries and merging HAVANA full length transcript annotation with the Ensembl gene build. In the last year, this process has been extended to include 4711 HAVANA pseudogenes and will be more regularly updated in future to incorporate additional validated annotation from GENCODE.

One additional geneset development is that the canonical transcripts are now defined for all genes and for all species. The canonical transcript is defined as either the

longest CDS, if the gene has translated transcripts, or the longest cDNA. Should a transcript already regarded as canonical not be selected using the above rules, there is support for storing this information in the Ensembl database.

Multiple alignments for comparative genomics

The genome-wide Ensembl comparative genomics pipeline has changed significantly over 2008, and is now based on the Enredo-Pecan-Ortheus pipeline (EPO). These are a set of three programs which feed into each other. The Enredo programme (28) takes a set of genomes and creates a segmentation graph across all the genomes to extract a set of colinear homologous segments. Unlike the algorithms Ensembl has used previously, Enredo handles lineage specific duplications (for example, a duplication on the primate lineage giving rise to two copies of a series of genes in primates compared to other mammals). These colinear segments are then handed onto Pecan, a consistency based multiple aligner, which provides a highly accurate alignment of the homologous regions. Using an assessment based on ancestral repeats, Enredo + Pecan outperforms other combinations of alignment programs in mammals. Finally, the ancestral sequence reconstruction programme, Ortheus (29), generates accurate ancestral sequences across each region. Ortheus uses a branch transducer model, a type of HMM, to call deletion and insertion events, providing a realistic model under which it can infer the ancestral sequence. Figure 2 shows the results of GERP (30) analysis of constraint across different feature types found in Ensembl, showing a sharp distinction between coding exons and ancestral repeats, with regulatory regions showing a intermediate level of constraint. Ensembl release 49 (March 2008) saw the first set of EPO alignments on a set of seven mammals. In release 50 (July 2008), this set of alignments was extended to include low-coverage genomes, creating a 23 mammals EPO alignment. A set of 4-way primate EPO alignments was also added containing human, chimp, orangutan, macaque. We plan to produce EPO multiple alignments in the teleost lineage in the future.

To create the 23 mammal EPO alignment, the methodology had to be extended to include low-coverage genomes. The assemblies of low-coverage genomes are too fragmented, creating too many breakpoints in the Enredo graph, to use Enredo directly. The Enredo graph was therefore built using high-coverage genomes only. Low-coverage genomes were then mapped on the colinear regions using pairwise alignments to the human genome. For each low-coverage genome, the segments defined by the pairwise alignments were linked with stretches of N's to facilitate the process of building the final multiple sequence alignment. After the alignment has been obtained, the stretches of N's were removed.

As well as providing alignments of genome sequence, the Ensembl comparative genome analysis pipelines also generate gene trees and orthology/paralogy prediction across all Ensembl genomes. A full description of the pipeline including its close collaboration between the curated resource Treefam (31) is forthcoming (Vilella, A. *et. al.*, submitted). A significant recent change (release 50) has been the calculation of site-wise dN/dS values in our gene trees using the SLR programme (sitewise likelihood ratio estimation of selection) (32). These values allow us to detect positions in the alignments that are under different evolutionary pressure.

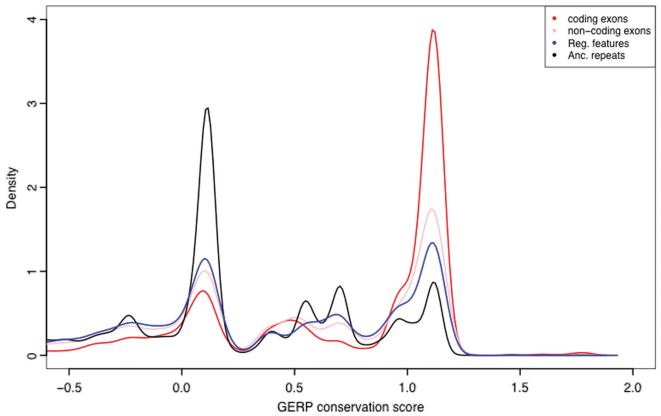

Figure 2. Figure shows a smoothed density plot of the GERP conservation scores (30) calculated from the 9-way EPO mammalian genome alignment corresponding to human chromosome X (Ensembl release 51). Four different types of genomic features are plotted: coding exons (red), non-coding exons (pink), regulatory features (blue) and ancestral repeats (black). A GERP score of 0 indicates no evidence of selective constraint, whereas high GERP scores shows evidence of selective constraint. Non-coding exons include all non-coding positions of protein-coding genes. Regulatory features include all regulatory features defined by the Ensembl regulatory build ('Gene Associated', 'Non-Gene Associated', 'Promoter Associated' and 'Unclassified'). Ancestral repeats include MER type II transposons only, as defined by RepeatMasker. Conserved features, such as exons and regulatory features, are clearly distinguished from repeats, a good indication of the quality of the EPO alignments.

Functional Genomics and Variation resources

The availability of genome wide functional data is one of the major changes in genomics in the last few years. Driven by involvement in analysis for the ENCODE project (9) and other international research consortia such as the EU FP6 funded HEROIC (High-throughput Epigenetic Regulatory Organisation In Chromatin) project, Ensembl has built up an infrastructure to support handling and display of this class of data (14). We have also recently participated in the creation of a genome-wide DNA methylation resource that has been incorporated into Ensembl (33,34). With the availability of next generation sequencing technology, array based ChIP-chip functional data is very rapidly giving way to sequence based ChIP-seq data. A major activity this year has been the development of a ChIP-seq analysis pipeline including a custom algorithm for the analysis of ChIP-seq data.

One of the characteristic features of the Ensembl project has been to go beyond presenting raw data aligned to the genome sequence by also presenting high quality consensus biological predictions, generated from automatic analysis pipelines developed to use the raw data as evidence. Examples are the Ensembl gene build pipeline generating

protein coding genesets and the Ensembl comparative analysis pipelines generating genetrees and orthology and paralog relationships. The Ensembl regulatory build is the latest such pipeline and provides automatic, evidence based annotation of potential regulatory regions within the human genome. The primary inputs are maps of open chromatin created by DNase I hypersensitivity mapping and covalent modifications of histone protein tails assayed by chromatin immunoprecipitation (ChIP). The first build was released in coordination with the ENCODE Pilot Project publication (9). Since the first release reported last year (14), we have updated the regulatory build three times, each time adding more data (35,36) and a more sophisticated analysis of the chromatin conformation and modification data. The build now consists of approximately 175 000 genomic regions defined from data collected from several cell types, including CD4 cells which make up the majority of the supporting data. Approximately 40 different histone modifications are now included and more than 2700 combinations of these factors form patterns associated with protein coding genes or their promoters allowing over 23 000 of the regulatory features to be classified as gene- or promoter-associated.

The rapid adoption of next generation sequencing technologies is also having a major impact on variation data in Ensembl. Whereas data continues to be imported from dbSNP, a major new source of computationally discovered variation data is from the processing of resequencing data. This second data source is growing rapidly in parallel with next generation sequencing technology. This year, Ensembl imported the data from three successive builds of dbSNP (127, 128, and 129). It has also incorporated resequencing-based SNPs from platypus and orangutan and as well as from the resequenced human genomes of Watson and Venter. The playpus SNPs were submitted to dbSNP and make up the largest set of SNPs for that species. The orangutan SNPs will be submitted in conjunction with the publication of that genome.

Within the variation database, we have increased support for copy number variation data and annotation of individual SNPs [e.g. with disease associations identified in genome-wide scans and with expression QTLs (37)]. The Ensembl variation group is synergistic with the European Genotype Archive (EGA http://www.ebi.ac.uk/ega/) and the 1000 Genomes Project (http://www.1000genomes.org/) data coordination centre groups at the EBI (European Bioinformatics Institute). The EGA was launched in the spring of 2008 and currently manages data from several projects including the Wellcome Trust Case Control Consortium (38) and other projects that are still in pre-publication status. The synergies between these projects will underpin the growth in variation data in Ensembl and the start of its functional annotation.

Outreach

Ensembl continues to make a substantial investment in training and user support. We regard this as critical not only to help users, but also evaluate the relevance of the data we provide and the easy of use of the services we provide. As discussed earlier, user engagement has been critical in developing the web site redesign. The Ensembl Outreach and Training group provides on-site courses on request and has run 102 workshops since May 2007, with an expanding effort in Asia (workshops in China, Malaysia and India), and a substantial presence in USA (20 workshops) and Europe (64 workshops). In addition to this, alongside standalone video tutorials, eLearning courses are now being developed and piloted within the EBI training platform (http://www.ebi.ac.uk/training/user/). Finally the new Ensembl blog (http://ensembl.blogspot.com/) provides updates on upcoming Ensembl training courses around the world.

FUTURE DIRECTIONS

The impact of next generation sequencing on genomics is beginning to be felt and a major focus for Ensembl is adapting to changes in data type and scale that will result. As discussed last year (14) the scale of data is a major challenge for many bioinformatics resources. For the variation team the immediate challenge is to present the variation landscape that will be uncovered by the 1000 Genomes Project, which is now running. The gene build team is starting to develop pipelines that use next generation sequencing transcriptome data. We can envisage such data being collected systematically for many different cell types and developmental stages, providing increasingly complete evidence for alternative splicing variants and functional annotation of the time and localization of their expression.

At present the focus for genome sequencing is discovery of variation, however as both experimental and computational techniques improve, it will become possible to sequence and assemble large genomes *de novo*. At this point it may become cost effective to sequence many more mammalian genomes. However, a major expansion of the number of genomes provided using Ensembl technology is already underway, in the form of the Ensembl Genomes project, (http://www.ensemblgenomes.org/), which will use Ensembl technology to provide a common interface to genomes across biology. Significant API and schema developments have already taken place to support this, including the ability to store several species in a single core database.

Finally, it is clear from our website performance monitoring that despite the performance improvements from our improved web-code, network latency effects will always reduce performance for users far from our servers. As a result we have been investing in mirror sites in parallel, to improve performance for users and provide redundancy. We have recently deployed a mirror site in China in collaboration with the Beijing Genomics Institute, Shenzhen (BGI-SZ). This site's primary service region is our users in and around China as the connections between the UK and China are relatively slow. We will shortly be deploying a full mirror to the US west coast and have also been investigating operating servers in commercially managed cloud compute facilities.

ACKNOWLEDGEMENTS

We acknowledge those researchers and organizations that have provided data to Ensembl prior to publication under the understandings of the Fort Lauderdale meeting discussing Community Resource Projects. We thank all our users of our website and other resources, and those who have provided useful feedback through our mailing list.

FUNDING

This work was supported by the Wellcome Trust [grant numbers WT062023]; the European Molecular Biology Laboratory (EMBL); the National Institutes of Health (NIH) National Human Genome Research Institute (NHGRI); the National Institutes of Health (NIH) National Institute of Allergy and Infectious Diseases (NIAID); the Biotechnology and Biological Sciences Research Council (BBSRC); the Medical Research Council (MRC); and the European Union. Funding for open access charge: The Wellcome Trust.

Conflict of interest statement. None declared.

REFERENCES

1. Karolchik,D., Kuhn,R.M., Baertsch,R., Barber,G.P., Clawson,H., Diekhans,M., Giardine,B., Harte,R.A., Hinrichs,A.S., Hsu,F. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.*, **36**, D773–D779.
2. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2008) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **36**, D13–D21.
3. Dowell,R.D., Jokerst,R.M., Day,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.
4. Jenkinson,A.M., Albrecht,M., Birney,E., Blankenburg,H., Down,T., Finn,R.D., Hermjakob,H., Hubbard,T.J., Jimenez,R.C., Jones,P. *et al.* (2008) Integrating biological data – the Distributed Annotation System. *BMC Bioinformatics*, **9(Suppl 8)**, S3.
5. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
6. The Star Consortium (2008) SNP and haplotype mapping for genetic analysis in the rat. *Nat. Genet.*, **40**, 560–566.
7. Twigger,S.N., Pruitt,K.D., Fernández-Suárez,X.M., Karolchik,D., Worley,K.C., Maglott,D.R., Brown,G., Weinstock,G., Gibbs,R.A., Kent,J. *et al.* (2008) What everybody should know about the rat genome and its online resources. *Nat. Genet.*, **40**, 523–527.
8. Aitman,T.J., Critser,J.K., Cuppen,E., Dominiczak,A., Fernandez-Suarez,X.M., Flint,J., Gauguier,D., Geurts,A.M., Gould,M., Harris,P.C. *et al.* (2008) Progress and prospects in rat genetics: a community view. *Nat. Genet.*, **40**, 516–522.
9. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
10. Gräf,S., Nielsen,F.G., Kurtz,S., Huynen,M.A., Birney,E., Stunnenberg,H. and Flicek,P. (2007) Optimized design and assessment of whole genome tiling arrays. *Bioinformatics*, **23**, i195–i204.
11. Warren,W.C., Hillier,L.W., Marshall Graves,J.A., Birney,E., Ponting,C.P., Grutzner,F., Belov,K., Miller,W., Clarke,L., Chinwalla,A.T. *et al.* (2008) Genome analysis of the platypus reveals unique signatures of evolution. *Nature*, **453**, 175–183.
12. Mikkelsen,T.S., Wakefield,M.J., Aken,B., Amemiya,C.T., Chang,J.L., Duke,S., Garber,M., Gentles,A.J., Goodstadt,L., Heger,A. *et al.* (2007) Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. *Nature*, **447**, 167–177.
13. Rhesus Macaque Genome Sequencing and Analysis Consortium. (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–234.
14. Flicek,P., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2008) Ensembl 2008. *Nucleic Acids Res.*, **36**, D707–D714.
15. Spudich,G., Fernandez-Suarez,X.M. and Birney,E. (2007) Genome browsing with Ensembl: a practical overview. *Briefings in functional genomics & proteomics*, **6**, 202–219.
16. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
17. Cunningham,F., Rios,D., Griffiths,M., Smith,J., Ning,Z., Cox,T., Flicek,P., Marin-Garcin,P., Herrero,J., Rogers,J. *et al.* (2006) TranscriptSNPView: a genome-wide catalog of mouse coding variation. *Nat. Genet.*, **38**, 853.
18. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F., Cutts,T. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
19. Wilming,L.G., Gilbert,J.G., Howe,K., Trevanion,S., Hubbard,T. and Harrow,J.L. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.
20. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
21. Denoeud,F., Kapranov,P., Ucla,C., Frankish,A., Castelo,R., Drenkow,J., Lagarde,J., Alioto,T., Manzano,C., Chrast,J. *et al.* (2007) Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. *Genome Res*, **17**, 746–759.
22. Harrow,J., Denoeud,F., Frankish,A., Reymond,A., Chen,C.K., Chrast,J., Lagarde,J., Gilbert,J.G., Storey,R., Swarbreck,D. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol*, **7(Suppl 1)**, S41–S49.
23. Guigó,R., Flicek,P., Abril,J.F., Reymond,A., Lagarde,J., Denoeud,F., Antonarakis,S., Ashburner,M., Bajic,V.B., Birney,B. *et al.* (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biology*, **7**, S2.
24. Tress,M.L., Martelli,P.L., Frankish,A., Reeves,G.A., Wesselink,J.J., Yeats,C., Olason,P.L., Albrecht,M., Hegyi,H., Giorgetti,A. *et al.* (2007) The implications of alternative splicing in the ENCODE protein complement. *Proc. Natl Acad. Sci. USA*, **104**, 5495–5500.
25. Clamp,M., Fry,B., Kamal,M., Xie,X., Cuff,J., Lin,M.F., Kellis,M., Lindblad-Toh,K. and Lander,E.S. (2007) Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl Acad. Sci. USA.*, **104**, 19428–19433.
26. The UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.
27. Bruford,E.A., Lush,M.J., Wright,M.W., Sneddon,T.P., Povey,S. and Birney,E. (2008) The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Res.*, **36**, D445–D448.
28. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: Genome-wide mammalian consistency based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
29. Paten,B., Herrero,J., Fitzgerald,S., Beal,K., Flicek,P., Holmes,I. and Birney,E. (2008) Genome-wide nucleotide level mammalian ancestor reconstruction. *Genome Res.*, **18**, 1829–1843.
30. Cooper,G.M., Stone,E.A., Asimenos,G., Program,N.C.S., Green,E.D., Batzoglou,S. and Sidow,A. (2005) Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res.*, **15**, 901–913.
31. Ruan,J., Li,H., Chen,Z., Coghlan,A., Coin,L.J., Guo,Y., Heriche,J.K., Hu,Y., Kristiansen,K., Li,R. *et al.* (2008) TreeFam: 2008 Update. *Nucleic Acids Res.*, **36**, D735–D740.
32. Massingham,T. and Goldman,N. (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics*, **169**, 1753–1762.
33. Down,T.A., Rakyan,V.K., Turner,D.J., Flicek,P., Li,H., Kulesha,E., Graf,S., Johnson,N., Herrero,J., Tomazou,E.M. *et al.* (2008) A Bayesian deconvolution strategy for

immunoprecipitation-based DNA methylome analysis. *Nat. Biotechnol.*, **26**, 779–785.
34. Rakyan,V., Down,T., Thorne,N., Flicek,P., Kulesha,E., Graf,S., Tomazou,E., Backdahl,L., Johnson,N., Herberth,M. *et al.* (2008) An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res.*, **18**, 1518–1529.
35. Wang,Z., Zang,C., Rosenfeld,J.A., Schones,D.E., Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Peng,W., Zhang,M.Q. *et al.* (2008) Combinatorial patterns of histone acetylations and methylations in the human genome. *Nat. Genet.*, **40**, 897–903.
36. Barski,A., Cuddapah,S., Cui,K., Roh,T.Y., Schones,D.E., Wang,Z., Wei,G., Chepelev,I. and Zhao,K. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.
37. Stranger,B.E., Nica,A.C., Forrest,M.S., Dimas,A., Bird,C.P., Beazley,C., Ingle,C.E., Dunning,M., Flicek,P., Koller,D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.
38. Wellcome Trust case control consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.